ABSTRACT

Human Action Recognition (HAR) in videos is a popular field in computer vision, focusing on human activities in various situations. Accurate recognition of activities is challenging due to noisy environments, occlusions, and perspective dissimilarities. Skeleton-based HAR models have been developed to learn the appearance and temporal dynamics of body joints, but these are local characteristics and rely heavily on skeleton prediction. Deep learning models, such as Convolutional Neural Networks (CNNs), have been developed to learn a hierarchy of features and train them using supervised or unsupervised approaches. As a result, a Joint-pooled 3D-Deep convolutional Descriptor (JDD) was developed using the two-stream Convolutional 3D (C3D) model to learn body joint guidance and generate a video descriptor. Such video descriptors were later classified by the Support Vector Machine (SVM) classifier into different actions. However, it did not learn spatiotemporal dependencies among large-scale video sequences and it was time-consuming for the skeleton extraction process. Thus, this research aims to analyze the challenges in the CNN-based HAR models and provide new solutions by improving the CNN model for HAR from large-scale video sequences.

The first phase of the research proposes the Joints and Trajectory-pooled 3D-Deep convolutional Descriptor (JTDD) model, in which an optical flow is integrated with the two-stream bilinear model to enhance recognition accuracy. In this model, an optical flow or trajectory point between video frames is also extracted at the body joint positions as input to the JTDD for creating the video descriptors. Those video descriptors are then classified by the SVM classifier to recognize human activities.

The second phase of the research proposes the Joints and Trajectory-pooled 3D-Deep Positional Attention-based Bidirectional Recurrent convolutional Descriptor (JTPADBRD) model that integrates the Positional-Attention Bidirectional Recurrent Neural Network (PABRNN) model as a pooling layer into a two-stream C3D network to extract significant spatiotemporal features and increase the accuracy of recognizing individual activities. This layer can aggregate convolutional feature vector representations of each clip belonging to a single video. Also, the activations of fully connected layers and their spatiotemporal variances are aggregated to generate the final

video descriptor. This video descriptor is applied to the SVM to recognize the individual activities in video sequences.

The third phase of the research proposes the Joints and Trajectory-pooled 3D-Deep Positional Attention-based Hierarchical Bidirectional Recurrent Convolutional Descriptor model (JTDPAHBRD), which introduces a Positional Attention-based Hierarchical BRNN (PAHBRNN) to enhance the feature aggregation process. First, the entire video is segregated into multiple blocks and they are provided to the C3D network, which applies the PAHBRNN as a feature pooling layer. In PAHBRNN, the feature vectors related to the different parts of a human skeleton in a certain clip are hierarchically aggregated based on the position-aware guidance vector to create the video descriptor. Further, the SVM classifier is applied to classify the resultant video descriptor to recognize the person's actions.

The fourth phase of the research proposes the JTD-Geometric and PAHBRD model (JTDGPAHBRD), which explores the geometries of the graph structure, namely joints, edges, and surfaces, for HAR by considering skeletons as a sequence of graph joints. A new deep paradigm is proposed for learning discriminative interpretations of activities from primitive geometric information along with the trajectory coordinates. Also, new view conversion and temporal dropout layers are incorporated in the PAHBRNN to learn robust interpretations for HAR. Moreover, a fully connected layer is used to get the absolute video descriptor, which is later classified by the SVM classifier for recognizing human activities.

The fifth phase of the research proposes the Graph Convolutional Network (GCN) with the JTDGPAHBRD (JTDGPAHBRD-GCN) to create a video descriptor for HAR. The GCN can obtain complementary information, such as higher-level spatial-temporal features, between consecutive frames for enhancing end-to-end learning. Also, to improve feature representation ability, a search space with several adaptive graph components is created. Then, a sampling and computation-effective evolution scheme are applied to explore this space. Moreover, the resultant GCN provides the temporal dynamics of the skeleton pattern, which are fused with the geometric features of the skeleton body joints and trajectory coordinates from the JTDGPAHBRD to create a more effective video descriptor for HAR. Such obtained

descriptors are later classified by the SVM algorithm to recognize a variety of human actions. Finally, these different models are validated using the Penn Action dataset to measure their performance in recognizing human actions compared to the other models. The experimental results proved that the JTDGPAHBRD-GCN model on the Penn Action dataset achieved 99.82% accuracy for HAR compared to the other models.